Big Data Analytics Technical Project Proposal

Name : Arya Sasi
SUID : L00179434
Course : MSc in Big Data Analytics

Problem Description

In the vast field of healthcare, understanding how symptoms, demographics, and health indicators connect is a puzzle we aim to solve using the Comprehensive Disease Symptom and Patient Profile Dataset. We want to explore this dataset to learn more about how symptoms like fever, cough, and fatigue relate to a person's age, gender, and health stats. The goal is to find hidden patterns and unique profiles for different diseases. Whether you're a medical researcher, healthcare professional, or just curious about data, this collection offers valuable insights into health patterns. Uncover hidden trends, discover unique symptom profiles, and gain a deeper understanding of medical conditions. This dataset has the potential to revolutionize our comprehension of healthcare. From the dataset we can compare different symptoms for same disease for different peoples also. This dataset can be used by various stakeholders, including:

- Healthcare Professionals
- Medical Researchers
- Healthcare Technology Companies

Dataset

The dataset for this project will be obtained from kaggle and consists of information related comprehensive compilation of symptoms and patient profiles for a wide range of diseases. The dataset is organized as a CSV file and includes a thorough collection of attributes. The dataset mainly contain below mentioned columns:

- Disease: The name of the disease or medical condition.
- Fever: Indicates whether the patient has a fever (Yes/No).
- Cough: Indicates whether the patient has a cough (Yes/No).
- Fatigue: Indicates whether the patient experiences fatigue (Yes/No).
- Difficulty Breathing: Indicates whether the patient has difficulty breathing (Yes/No).
- Age: The age of the patient in years.
- Gender: The gender of the patient (Male/Female).

- Blood Pressure: The blood pressure level of the patient (Normal/High).
- Cholesterol Level: The cholesterol level of the patient (Normal/High).

Goal

The main goals of this project are to answer the below mentioned Questions:

1. For which disease the outcome Variable is more positive.
2. For which gender the cholesterol and Blood Pressure level is high.
3. Which Disease have the most positive and negative Outcome Variable.

Data Analytic Pipeline of Proposed Solution

- In the first step, called Data Collection (shown in Figure 1, we gather information about different diseases from the CSV file. This involves selecting data related to patients with various symptoms. It's a crucial starting point for our analysis.
- After that, this data would undergo pre-processing using methods like missing value check to remove noisy data from the data.
- The data transformation process is essential for getting the data into the correct format before feeding it into machine learning models. This step is crucial to ensure accurate predictions.

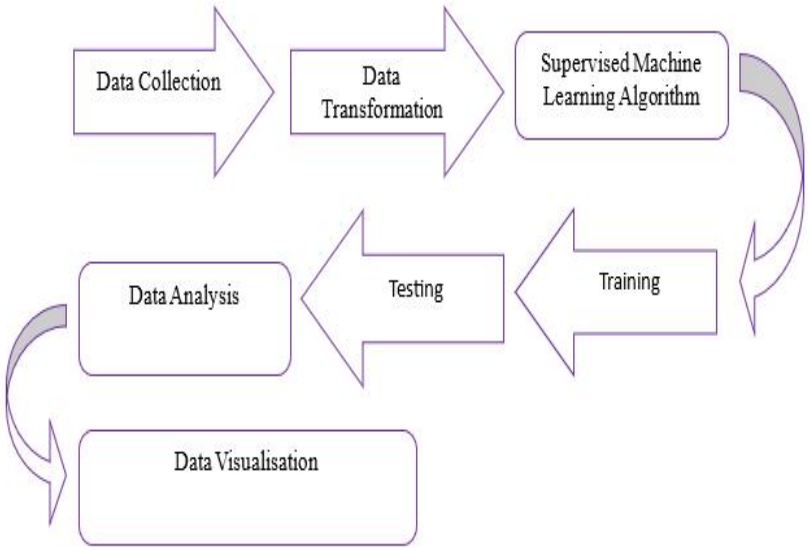

Figure 1: Proposed Methodology

References

[1] Kaggle

Link 1: Disease Symptoms and Patient Profile Dataset (kaggle.com)

Link2:https://www.mayoclinic.org/diseases-conditions/infectious-diseases/symptoms-causes/syc-20351